Partiel de modélisation statistique, première partie

M1 Mathématiques et Intelligence Artificielle, Université Paris-Saclay
2024-2025

Mardi 5 novembre 2024
13h30 – 16h30 (tiers-temps 13h30 – 17h30)

Une feuille A4 recto manuscrite est autorisée en tant que support. La calculatrice est autorisée.

Avant de commencer :

- *Le sujet comporte trois exercices. Il est demandé de les traiter dans l'ordre sur la copie.*
- *Des résultats pourront être admis d'une question sur l'autre, à la condition de l'écrire clairement.*
- *Une attention particulière sera portée aux questions d'interprétation, à la rigueur et la concision de la rédaction.*
- *Un barème indicatif est donné :*

  Exercice 1 : 2 points,

  Exercice 2 : 8 points,

  Exercice 3 : 10 points.

Exercice 1 – Pour se mettre dans le bain

Soit $n \geq 1$ et $X_1, \ldots, X_{2n}$ des variables réelles i.i.d. de loi $\mathcal{N}(0,1)$. On pose

$$U = \frac{1}{2}\sum_{i=1}^{n}(X_{2i} - X_{2i-1})^2 \quad \text{et} \quad V = \sum_{i=1}^{2n} X_i \,.$$

1. Quelle est la loi de V ? *Solution. Par transformation linéaire, V est gaussienne. Il est clair qu'elle est de moyenne nulle, de variance $2n$.*

2. Montrer que U et V sont indépendantes. On pourra poser Z le vecteur suivant :

$$Z := (X_2 - X_1, X_4 - X_3, \ldots, X_{2n} - X_{2n-1}, V)^T,$$

   Solution. Le vecteur Z est gaussien par transformation linéaire d'après le cours. Chaque terme de covariance impliquant V et un $X_{2i} - X_{2i-1}$ vaut exactement $1 - 1 = 0$. Et tous les autres termes non diagonaux sont nuls car ils impliquent des variables distinctes. La matrice de covariance de Z est donc $\mathrm{diag}(2, \ldots, 2, 2n)$*. Par transfomation, U est indépendante de V.*

3. Donner la loi de U en la justifiant soigneusement. *Solution. U est la norme au carré d'un vecteur de dimension n dont les coordonnées sont i.i.d. de loi $\frac{1}{\sqrt{2}}\mathcal{N}(0,2) = \mathcal{N}(0,1)$ (l'égalité précédente est une égalité en loi). Par définition de la loi du χ^2 on a donc $U \sim \chi^2(n)$.*

Exercice 2 – Durée de vie d'une pièce industrielle

Une entreprise de fabrication automobile cherche à expliquer et prédire la durée de vie d'une pièce de sa fabrication, en fonction de variables d'intérêt. Cette pièce, en alliage d'aluminium, est située dans le moteur des véhicules produits par l'entreprise. C'est un parrallélépipède rectangle dont la base est de surface constante, mais d'épaisseur (ou hauteur) variable. Elle est de densité variable selon l'alliage d'aluminium utilisé. On dispose d'un jeu de données de $n = 71$ observations dont les variables sont les suivantes :

- `duree_vie` : durée de vie de la pièce, définie comme le nombre de km parcourus au compteur du véhicule avant de devoir changer la pièce ;
- `epaisseur` : épaisseur de la pièce, en cm ;
- `poids` : poids de la pièce, en hg (1 hg = 10^2 g) ;
- `finition` : en fin de fabrication de la pièce en aluminium, une finition lui est appliquée. On distingue trois types de finitions, notés A, B et C.

Dans cet exercice, tous les modèles linéaires rencontrés sont supposés gaussiens. On donne ci-dessous un aperçu du début du jeu de données ainsi qu'un résumé de celui-ci.

duree_vie <dbl>	epaisseur <dbl>	poids <dbl>	finition <chr>
120913.10	6.265404	7.116102	C
139490.76	4.198606	3.115564	B
108487.94	5.286604	2.315019	A
98329.29	2.989635	1.562986	A
115342.99	5.635362	3.820599	C
97965.86	2.804801	1.294772	C

Figure 1: Aperçu du jeu de données

```
   duree_vie       epaisseur         poids          finition
Min.   : 76750   Min.   :1.889   Min.   :0.3094   Length:71
1st Qu.:102958   1st Qu.:3.183   1st Qu.:0.4542   Class :character
Median :118248   Median :3.997   Median :0.7420   Mode  :character
Mean   :118682   Mean   :3.935   Mean   :0.9292
3rd Qu.:131886   3rd Qu.:4.726   3rd Qu.:1.1978
Max.   :161501   Max.   :6.444   Max.   :3.1897
```

Figure 2: Résumé du jeu de données

```
Call:
lm(formula = duree_vie ~ ., data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-26797.8  -6220.6   -128.4   5958.2  28778.1

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 100976.0     5344.9  18.892  < 2e-16
epaisseur      162.6     1299.7   0.125    0.901
poids         3266.2      522.2   6.254 3.35e-08
finitionB    20831.8     3812.8   5.464 7.64e-07
finitionC    -3387.5     3494.3  -0.969    0.336
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10980 on [redacted] degrees of freedom
Multiple R-squared:  0.7004,    Adjusted R-squared:  0.6822
F-statistic: 38.57 on [redacted] and [redacted] DF,  p-value: < 2.2e-16
```

Figure 3: Résultat de la première régression

1. On commence par une régression linéaire classique dont le résultat est présenté sur la figure 3. Attention, certains éléments sont effacés.

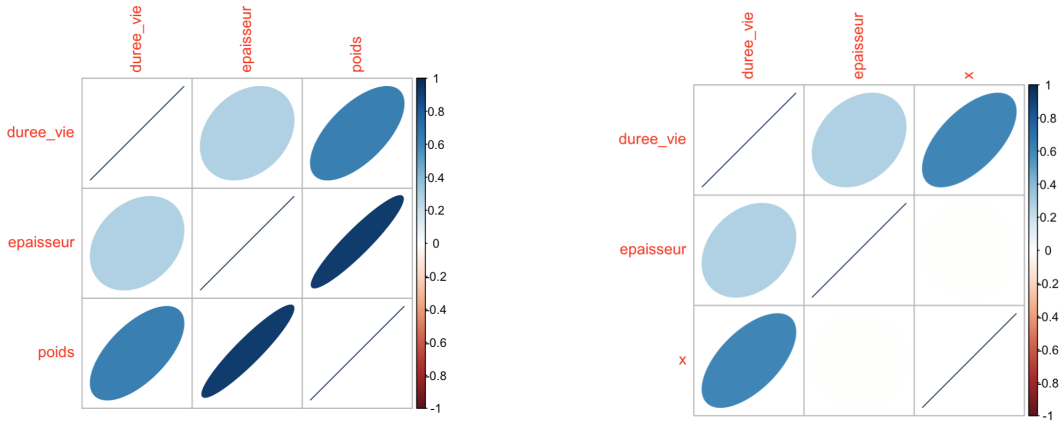
   Au vu de ces résultats, quelles sont les éléments expliquant significativement la durée de vie de la pièce ? Dans quel sens sont-ils corrélés à la variable réponse ? On appuiera la réponse sur des arguments quantititatifs. *Solution. c'est une question très gentille ; les facteurs expliquant significativement la durée de vie de la pièce sont a priori sont poids, et le fait ou non qu'elle ait reçu la finition B. Ce sont les seuls coefficients pour lesquels la p-valeur du test de Student associé est très faible ($\leq 10^{-6}$). Ces deux facteurs sont positivement corrélés à la variable réponse. Les autres ne semblent pas significatifs.*

2. Soucieux d'étudier les données plus en détail, on mène une analyse bivariée des variables quantitatives et on obtient les corrélations empiriques présentées sur la figure 4(a).

   2.(a). Que dire quant à la redondance des variables explicatives ? Donner une explication de cette éventuelle redondance. On souhaite remplacer la variable `poids` par une variable notée `x`. Proposer une expression pertinente de la variable `x` pour supprimer la redondance dans les données. *Solution. Bien sûr, le poids et l'épaisseur sont intrinsèquement corrélés : la pièce étant de surface constante, il y a relation linéaire directe entre le poids et l'épaisseur. Au vu de l'énoncé, les alliages étant de densité différentes, c'est la densité qui est intéressante. On crée donc cette variable* $x = \frac{poids}{épaisseur}$ *qui supprimera la redondance observée ci-dessus.*

   2.(b). Nous appliquons à notre jeu de données la modification proposée en question 2.(a). : la nouvelle variable `x` remplace désormais la variable `poids`. Commenter la figure 4(b).

3. Sur la figure 5, on présente le résultat de la régression linéaire pour ce jeu de données modifié. Compléter les trois dernières lignes du résultat affiché dans la figure 5 en expliquant votre raisonnement. *Solution. Grâce à* X^TX *(où à la figure 2 si on observe bien), on lit* $n = 71$. *De plus, ici* $p = 1 + 1 + 1 + 2 = 5$ *en comptant l'intercept et la variable qualititative. On en déduit que c'est un* $n - p = 66$ *qui manque pour l'estimée de* σ^2, *puis* $p - 1 = 4$ *et* $n - p = 66$ *à la dernière*



(a) Corrélations empiriques

(b) Nouvelles corrélations empiriques

Figure 4

```
Residuals:
     Min       1Q   Median       3Q      Max
-25561.9  -5573.7   -713.6   4851.0  28125.7

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    86464       5572  15.517  < 2e-16
epaisseur       3508                    0.00387
finitionB      20366       3704   5.498 6.69e-07
finitionC      -2867       3392  -0.845  0.40114
x              14192       2098   6.765 4.23e-09
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10640 on    degrees of freedom
Multiple R-squared:  0.7182,    Adjusted R-squared:
F-statistic: 42.06 on    and    DF,  p-value: < 2.2e-16
```

Figure 5: Résultats de la régression pour le jeu de données modifié

ligne. Le R^2 ajusté vaut d'après le cours : $R^2_a = 1 - \frac{n-1}{n-p}(1-R^2)$. Application numérique : $R^2_a = 1 - \frac{70}{66}(1-0.7182)) \sim 0.7011$.

4. Ce deuxième modèle (avec `x`) est-il plus pertinent que le premier (avec `poids`) pour expliquer les données ? Donner deux indicateurs quantitatifs pour justifier votre réponse. *Solution. Oui ; le R^2 et le R^2 ajusté sont meilleurs, et la F statistique est plus haute à mêmes degrés de libertés : on rejette encore plus la nullité du modèle.*

5. La figure 6 donne des extraits de code. Retrouver les éléments cachés à la ligne `epaisseur` de la figure 5. On ne demande pas de retrouver les étoiles. *Solution. Ce sont l'estimée de l'écart-type du coefficient correspondant à* `epaisseur`*, ainsi que la statistique de Student correspondante. L'estimateur sans biais de l'écart-type du coefficient θ_i est donné par $\hat{\sigma}\sqrt{[(X^TX)^{-1}]_{i,i}}$, l'application numérique donne $10640 \times \sqrt{0.012117616} \sim 1171$ et la t-value est juste le quotient de l'estimée par son écart-type, autrement $t = \frac{3508}{1171} \sim 2.9957$.*

6. On s'intéresse à la variable `finition`. Pour $f \in \{A, B, C\}$, on note (f) le groupe des pièces telles que `finition` $= f$.

   6.(a). Les groupes (A) et (B) sont-ils statistiquement équivalents ? *Solution. On lit dans la figure 5 les résultats du test de nullité du paramètre θ_B avec A comme référence. La p-valeur est très très faible : les groupes ne sont pas du tout équivalent statistiquement.*

```r
X = as.matrix(cbind(rep(1,n),epaisseur,finition=='B',finition=='C',x))
colnames(X) <- NULL # on enlève le nom des colonnes
t(X)%*%X
```

```
          [,1]       [,2]     [,3]      [,4]      [,5]
[1,]  71.00000  279.38630 23.00000  34.00000  65.97137
[2,] 279.38630 1184.43593 96.23539 131.34674 259.54139
[3,]  23.00000   96.23539 23.00000   0.00000  26.96826
[4,]  34.00000  131.34674  0.00000  34.00000  26.55636
[5,]  65.97137  259.54139 26.96826  26.55636  89.22035
```

```r
solve(t(X)%*%X)
```

```
            [,1]         [,2]        [,3]         [,4]         [,5]
[1,]  0.27406085 -0.045651410 -0.03859205 -0.068088863 -0.037914974
[2,] -0.04565141  0.012117616 -0.00612221 -0.001874548  0.000914027
[3,] -0.03859205 -0.006122210  0.12111580  0.071188752 -0.011453203
[4,] -0.06808886 -0.001874548  0.07118875  0.101582443  0.004045512
[5,] -0.03791497  0.000914027 -0.01145320  0.004045512  0.038842193
```

Figure 6: Extraits de code

6.(b). Les groupes (A) et (C) sont-ils statistiquement équivalents ? *Solution. Même chose ici, mais la p-valeur est de 0.4 environ : on ne rejette pas l'hypothèse nulle, les groupes sont donc équivalents statistiquement.*

6.(c). Les groupes (B) et (C) sont-ils statistiquement équivalents ? Justifier soigneusement votre réponse. *Solution. En toute rigueur, il faudrait réaliser le test avec B ou C comem référence. Cependant, on voit que l'estimée de θ_B est très positive, celle de θ_C est négative, et comme (A) et (B) sont déjà distinguables, (B) et (C) le sont aussi.*

7. Combien y a-t-il de pièces de finition A dans l'échantillon ? *Solution. Il faut lire dans X^TX que $n_B = 23, n_C = 34$. On en déduit $n_A = 71 - 23 - 34 = 14$.*

8. L'ingénieur auquel vous présentez les résultats vous déclare : "au vu de cette étude statistique, le fait d'appliquer la finition B augmente la durée de vie de la pièce. Suggérez-vous de n'utiliser plus que la finition B dans notre processus de fabrication ? "

   Que répondre à cette question au vu de l'étude réalisée ? Argumentez. Quelle nouvelle étude recommanderiez-vous éventuellement afin de vous donner plus d'éléments pour répondre à cette question ?

   Solution. Nous touchons ici au phénomène de causalité (déjà rencontré en cours). Ce que l'on a établi, c'est que le fait d'appliquer la finition B expliquait (au sens de la corrélation) une durée de vie supérieure. Mais cette corrélation n'est pas nécessairement causalité. Par exemple il se peut que les pièces n'éient pas été séléctionnées à l'aveugle (premier souci) ou, pire encore, que dans le processus de fabrication la finition B soit appliquée seulement à des pièces déjà plus résistantes. Il faudrait faire une étude ou l'on fabrique des pièces et où le choix de la finition est fait aléatoirement (et indépendamment de tout le reste). A ce moment là, on pourra mettre en lumière un effet causal. C'est ce qu'on appelle un test randomisé.

Exercice 3 – Minimisation de l'erreur de prédiction

Dans cet exercice, on s'intéresse à l'erreur de prédiction dans le modèle linéaire. Dans tous les modèles rencontrés, on fera les hypothèses classiques du modèle linéaire, non supposé gaussien, et on note σ^2 la variance du bruit.

Première partie : cas à une variable explicative.

On cherche à apprendre le modèle linéaire suivant pour prédire le réel Y_i en fonction d'une variable explicative réelle z_i :

$$\forall 1 \leq i \leq n, \quad Y_i = \beta_0 + \beta_1 z_i + \varepsilon_i \,.$$

On notera $\overline{y}$ et $\overline{z}$ les moyennes empiriques de $z = (z_1, \ldots, z_n)^T$ et de $Y = (Y_1, \ldots, Y_n)^T$ (attention, elles dépendent de n). On suppose que le modèle est identifiable et on note $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1)$ l'estimateur des moindres carrés de β.

Sous le même modèle, on observe une nouvelle valeur z_{n+1} de la variable explicative et on cherche à prédire la variable réponse Y_{n+1} avec l'estimateur

$$\hat{y}_{n+1} := \hat{\beta}_0 + \hat{\beta}_1 z_{n+1} = x_{n+1}^T \hat{\beta}, \quad \text{où } x_{n+1} := \begin{bmatrix} 1 \\ z_{n+1} \end{bmatrix} .$$

L'erreur de prédiction est définie par :

$$\mathrm{err}(z_{n+1}) := \mathbb{E}\left[(Y_{n+1} - \hat{y}_{n+1})^2\right] .$$

Notons que dans cette espérance, l'aléa vient du bruit ε_{n+1} dans Y_{n+1} ainsi que de l'aléa dans le $\hat{\beta}$.

1. Montrer que si S est un vecteur aléatoire de $\mathbb{R}^m$ de matrice de covariance (finie) C, alors pour tout vecteur $u \in \mathbb{R}^m$, $\mathrm{Var}(u^T S) = u^T C u$. *Solution. On peut développer pour voir que*

$$\begin{aligned} (u^T S) &= \left(\sum_{i=1}^m u_i S_i \right) \\ &= \sum_{1 \leq i,j \leq m} (u_i S_i, u_j S_j) \\ &= \sum_{1 \leq i,j \leq m} u_i u_i C_{i,j} = u^T C u \,. \end{aligned}$$

2. Que vaut $\mathbb{E}\left[Y_{n+1} - \hat{y}_{n+1}\right]$? *Solution. Elle est égale à $\mathbb{E}\left[x_{n+1}^T(\beta - \hat{\beta}) + \varepsilon_{n+1}\right] = x_{n+1}^T 0 + 0 = 0$ par nullité du biais de $\hat{\beta}$ et d'après les hypothèses sur les ε_i.*

3. Ecrire la matrice X du plan d'expérience dans ce modèle (on mettra l'intercept dans la première colonne), et montrer que

$$(X^T X)^{-1} = \frac{1}{n v(z)} \begin{bmatrix} \frac{1}{n}\sum_{i=1}^n z_i^2 & -\overline{z} \\ -\overline{z} & 1 \end{bmatrix},$$

avec $v(z)$ la variance empirique de z. *Solution. Dans notre cas on a $X = \begin{bmatrix} \mathbf{1}_n \mid z \end{bmatrix} \in \mathbb{R}^{n \times 2}$, et $X^T X = n \begin{bmatrix} 1 & \overline{z} \\ \overline{z} & \frac{1}{n}\sum_{i=1}^n z_i^2 \end{bmatrix}$ de déterminant $n v(z)$. Il vient donc $(X^T X)^{-1} = \frac{1}{n v(z)} \begin{bmatrix} \frac{1}{n}\sum_{i=1}^n z_i^2 & -\overline{z} \\ -\overline{z} & 1 \end{bmatrix}$.*

4. En utilisant les résultats des questions 1, 2 et 3, montrer que

$$\mathrm{err}(z_{n+1}) = \sigma^2 \left(1 + \frac{1}{n} + \frac{(z_{n+1} - \overline{z})^2}{\sum_{i=1}^n (z_i - \overline{z})^2} \right) .$$

Solution. Comme la variable dont on prend le moment d'ordre deux est centré d'après la question 2, on calcule en fait une variance.

$$\begin{aligned} \mathrm{err}(z_{n+1}) &= \mathrm{Var}(\varepsilon_{n+1} + x_{n+1}^T(\beta - \hat{\beta})) \\ &= \sigma^2 + x_{n+1}^T (\sigma^2 (X^T X)^{-1}) x_{n+1} \,. \end{aligned}$$

D'après la question 3 on a $(X^TX)^{-1} = \frac{1}{nv(z)}\begin{bmatrix}\frac{1}{n}\sum_{i=1}^n z_i^2 & -\overline{z}\\ -\overline{z} & 1\end{bmatrix}$ *avec* $v(z)$ *variance empirique de* z*. Cela donne*

$$\begin{aligned}\mathrm{err}(z_{n+1}) &= \sigma^2 + x_{n+1}^T(\sigma^2(X^TX)^{-1})x_{n+1}\\ &= \sigma^2 + \frac{\sigma^2}{nv(z)}\left(\frac{1}{n}\sum_{i=1}^n z_i^2 - 2\overline{z}z_{n+1} + z_{n+1}^2\right)\\ &= \sigma^2 + \frac{\sigma^2}{nv(z)}\left(v(z) + \overline{z}^2 - 2\overline{z}z_{n+1} + z_{n+1}^2\right)\\ &= \sigma^2\left(1 + \frac{1}{n} + \frac{(z_{n+1}-\overline{z})^2}{nv(z)}\right).\end{aligned}$$

5. Pour quelle(s) valeur(s) de z_{n+1} l'erreur de prédiction est-elle minimale ? Interpréter ce résultat. *Solution. L'erreur est minimale en* $z_{n+1} = \overline{z}$ *et vaut* $\sigma^2 + \sigma^2/n$*. On est toujours meilleur pour prédire lorsqu'on observe le barycentre de ce qu'on a déjà observé. En effet, intuitivement, c'est le point le plus typique.*

6. Quelle est la limite de l'erreur de prédiction minimale lorsque $n \to \infty$? Interpréter cette valeur limite. *Solution. Cette erreur tend vers* σ^2 *lorsque* n *est grand, c'est l'erreur minimale de prédiction qu'on puisse faire, car la variable* Y_{n+1} *est bruitée avec un bruit indépendant et de variance* σ^2 *: on ne peut pas faire mieux.*

Deuxième partie : cas général.

Nous considérons cette fois-ci le cas général à plusieurs covaraibles. Ici, pour tout $1 \le i \le n$,

$$Y_i = \beta_0 + \beta_1 z_{1,i} + \ldots + \beta_p z_{p,i} + \varepsilon_i\,.$$

On note

$$Z := \left[\begin{array}{c|c|c|c} & & & \\ z_1 & \ldots & z_p & \\ & & & \end{array}\right] \in \mathbb{R}^{n\times p},$$

avec pour tout $1 \le \ell \le p$, $z_\ell := \begin{bmatrix} z_{\ell,1} & \ldots & z_{\ell,n}\end{bmatrix}^T \in \mathbb{R}^n$. On note $\overline{z}$ le vecteur des moyennes empiriques $\overline{z} := \begin{bmatrix}\overline{z_1} & \ldots & \overline{z_p}\end{bmatrix}^T \in \mathbb{R}^p$ (dont les entrées dépendent de n).

7. En notant $\beta = \begin{bmatrix}\beta_0 & \beta_1 & \ldots & \beta_p\end{bmatrix}^T$, écrire le modèle matriciellement sous la forme $Y = X\beta + \varepsilon$ en précisant X. *Solution. On l'a déjà fait,* X *s'écrit*

$$X = \left[\begin{array}{c|c|c|c} & & & \\ \mathbf{1}_n & z_1 & \ldots & z_p \\ & & & \end{array}\right] \in \mathbb{R}^{n\times(p+1)},$$

8. Ecrire la matrice X^TX sous forme de 4 blocs faisant intervenir Z, $\overline{z}$ et n. *Solution. Déjà fait en séance d'exercices. On a que* $X^TX = \begin{bmatrix} n & n\overline{z}^T \\ n\overline{z} & Z^TZ\end{bmatrix}$

Dans toute la suite, on suppose que le modèle est identifiable.

9. On donne la formule d'inversion matricielle par blocs : soit A une matrice inversible s'écrivant par blocs $A = \begin{bmatrix} T & U \\ V & W\end{bmatrix}$ avec T inversible. Alors $Q = W - VT^{-1}U$ est inversible et l'inverse de A est :

$$A^{-1} = \begin{bmatrix} T^{-1} + T^{-1}UQ^{-1}VT^{-1} & -T^{-1}UQ^{-1} \\ -Q^{-1}VT^{-1} & Q^{-1}\end{bmatrix}.$$

On note $\Gamma := \frac{1}{n} Z^T Z - \overline{z}\,\overline{z}^T$. Ecrire la matrice $(X^T X)^{-1}$ sous la forme d'une matrice par blocs en fonction de $n, \overline{z}$ et Γ^{-1}. *Solution. La formule donne directement $Q = n\Gamma$ et*

$$(X^T X)^{-1} = \frac{1}{n} \begin{bmatrix} 1 + \overline{z}^T \Gamma^{-1} \overline{z} & -(\Gamma^{-1}\overline{z})^T \\ -\Gamma^{-1}\overline{z} & \Gamma^{-1} \end{bmatrix}.$$

10. Comme dans la première partie, on observe désormais un vecteur

$$x_{n+1} = (1, z_{n+1}) = (1, z_{1,n+1}, \dots z_{\ell,n+1})$$

et l'on cherche à prédire la variable réponse Y_{n+1} avec l'estimateur $\hat{y}_{n+1} := x_{n+1}^T \hat{\beta}$. L'erreur de prédiction est définie comme précédemment.

Exprimer $\mathrm{err}(z_{n+1})$ en fonction de n, de σ^2, du vecteur $(z_{n+1} - \overline{z})$ et de la matrice Γ^{-1}. *Solution. On reprend les calculs de la question 3., et on a encore*

$$\begin{aligned} \mathrm{err}(z_{n+1}) &= \sigma^2 + x_{n+1}^T (\sigma^2 (X^T X)^{-1}) x_{n+1} \\ &= \sigma^2 + \frac{\sigma^2}{n} \left(1 + z^T \Gamma^{-1} z - 2 z_{n+1}^T \Gamma^{-1} \overline{z} + z_{n+1}^T \Gamma^{-1} z_{n+1}\right) \\ &= \sigma^2 \left(1 + \frac{1}{n} + (z_{n+1} - \overline{z})^T \Gamma^{-1} (z_{n+1} - \overline{z})\right). \end{aligned}$$

11. On admet dans cette question que Γ est symétrique définie positive. Pour quelle(s) valeur(s) de z_{n+1} l'erreur de prédiction est-elle minimale ? *Solution. Ben du coup c'est évident, il faut dire que Γ^{-1} est encore symétrique définie positive, et le minimum est donc atteint en $z_{n+1} = \overline{z}$. Le reste du temps, c'est au-dessus !*

12. (Question bonus). Montrer que $\Gamma = \frac{1}{n} Z^T Z - \overline{z}\,\overline{z}^T$ est bien symétrique définie positive. On pourra chercher à l'écrire sous la forme $\frac{1}{n} W W^T$ avec W une matrice bien choisie. *Solution. Symétrie évidente. Ensuite, l'idée c'est de voir qu'elle ressemble fortement à une matrice de covariance. C'est même carrément une matrice de covariance empirique. Si on note $\tilde{Z}$ la matrice*

$$\tilde{Z} := \left[\begin{array}{c|c|c} z_1 - \overline{z_1}\mathbf{1}_n & \dots & z_p - \overline{z_p}\mathbf{1}_n \end{array}\right] \in \mathbb{R}^{n \times p},$$

on a que $\Gamma = \frac{1}{n} \tilde{Z}^T \tilde{Z}$. On a donc la positivité qui en découle. On peut ensuite évoquer le résultat d'inversion par blocs qui implque que Γ est inversible, ou remarquer que $\tilde{Z}$ est injective. En effet, si $u \in \mathrm{Ker}\tilde{Z}$, alors $u_1 z_1 + \dots + u_p z_p = \left(\sum_{\ell=1}^p u_\ell \overline{z_\ell}\right) \mathbf{1}_n$, et comme X est de rang plein par hypothèse, $u = 0$.

Fin du sujet.